3. cvičení

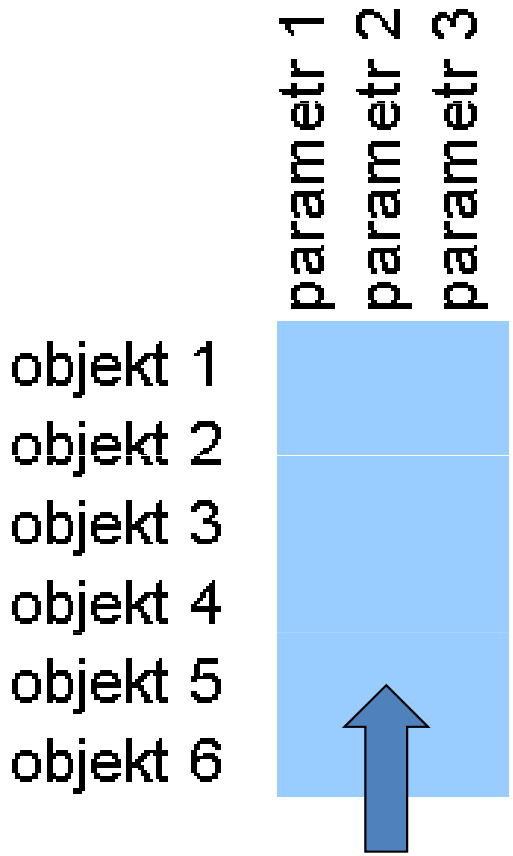
11.11.2014

Úvod do vícerozměrných metod I.

- **Vícerozměrné metody**: Název vícerozměrné vychází z typu vstupních dat, tato data jsou tvořena jednotlivými objekty (i.e. klienti) a každý z nich je charakterizován svými parametry (věk, příjem atd.) a každý z těchto parametrů můžeme považovat za jeden rozměr objektu.
- **Maticová algebra**: Základem práce s daty a výpočtů vícerozměrných metod je maticová algebra, matice tvoří jak vstupní, tak výstupní data a probíhají na nich výpočty.
- **NxP matice**: N objektů s p parametry pak vytváří tzv. NxP matici, která je prvním typem vstupu dat do vícerozměrných analýz.
- **Asociační matice**: Na základě těchto matic jsou počítány matice asociační na nichž pak probíhají další výpočty, jde o čtvercové matice obsahující informace o podobnosti nebo rozdílnosti (tzv. **metriky**) buď objektů (Q mode analýza) nebo parametrů (R mode analýza).Měřítko podobnosti se liší podle použité metody a typu dat, některé metody umožňují použití uživatelských metrik.

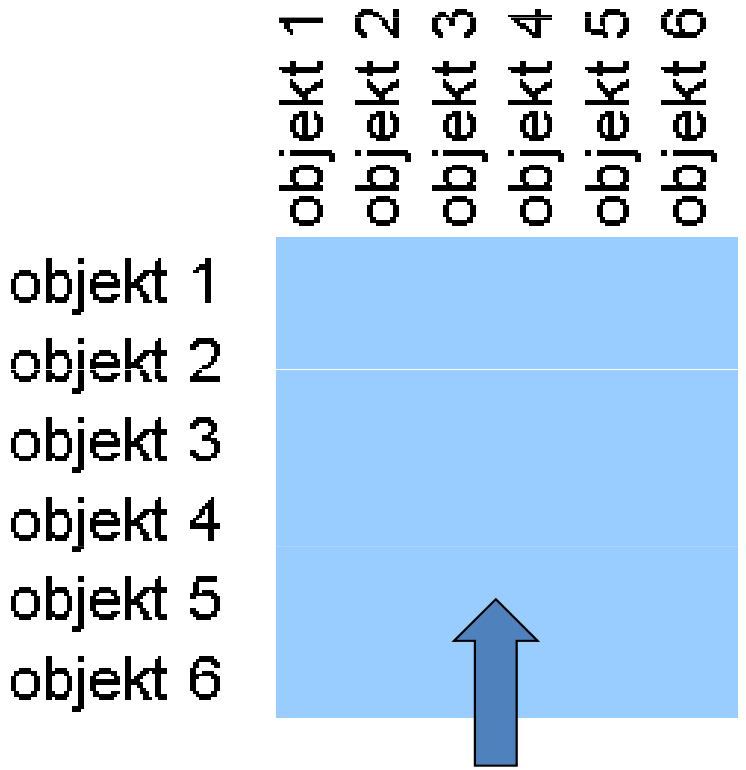
Vstupní matice vícerozměrných analýz

NxP MATICE



Hodnoty parametrů pro jednotlivé objekty

ASOCIAČNÍ MATICE



Korelace, kovariance, vzdálenost, podobnost

Úvod do vícerozměrných metod II.

SHLUKOVÁ ANALÝZA

- vytváření shluků objektů na základě jejich podobnosti
- identifikace typů objektů

ORDINAČNÍ METODY

- zjednodušení vícerozměrného problému do menšího počtu rozměrů
- principem je tvorba nových rozměrů, které lépe vyčerpávají variabilitu dat

Měření vzdálenosti objektů

Euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} (x_{ik} - x_{jk})^2}$$

Vážená euklidovská vzdálenost

$$d_{ij} = \sqrt{\sum_{k=1}^{p} w_k^2 (x_{ik} - x_{jk})^2}$$

i,j – označení objektů
d_{ij} – vzdálenost objektů i a j
p – počet parametrů
k – k-tý parametr
w_k – váha parametru k

Minkowski (power distance)

$$d_{ij} = \sqrt[\frac{1}{\lambda}]{\sum_{k=1}^{p} \left| x_{ik} - x_{jk} \right|^{\lambda}}$$

- - celé číslo
- =1 Manhattan (city block)
- = 2 Euklidovská vzdálenost

Chebychev

$$d_{ij} = \max \left| x_{ik} - x_{jk} \right|$$

IBA MU

Měření podobnosti objektů

Binární koeficienty podobnosti

Objekt 2 \ Objekt 1	1	0
1	a	b
0	c	d

a, b, c, d = počet případů, kdy souhlasí binární charakteristika objektu 1 a 2
a+b+c+d=p

Symetrické binární koficienty - není rozdíl mezi případem 1-1 a 0-0

Simple matching coefficient

$$S(x_1 x_2) = \frac{a+d}{p}$$

Hamman, Yule coefficient, Pearson's ↗ (phi) a další koeficienty

Asymetrické binární koeficienty – odstranění double zero

Jaccard`s coefficient

$$S(x_1 x_2) = \frac{a}{a+b+c}$$

Sorensen`s coefficient

$$S(x_1 x_2) = \frac{2a}{2a+b+c}$$

Řada dalších koeficientů dávajících různou váhu jednotlivým kombinacím parametrů

Kvantitativní koeficienty

Obdoby binárních koeficientů pro více parametrů než 0/1

Simple matching coefficient pro více parametrů

$$S(x_1 x_2) = \frac{souhlas}{p}$$

p=počet parametrů

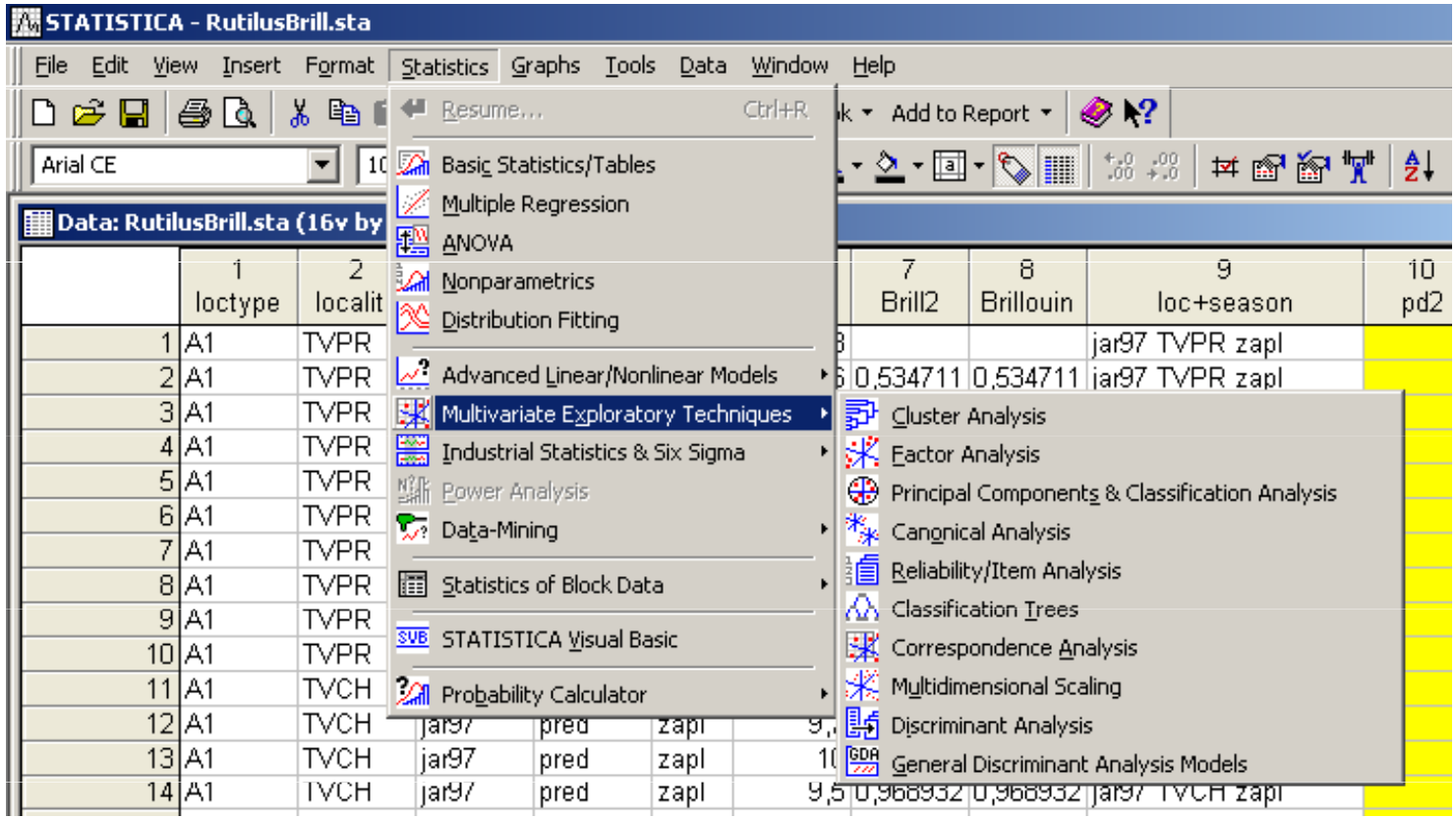
Gowerův koeficient

Zahrnutí podobnosti podle různých typů parametrů – binární, kvalitativní a semikvantitativní i kvantitativní (odlišný výpočet pro jednotlivé typy). Celkový součet podobností je podělen počtem parametrů. Může zahrnovat podmínku nepočítat s chybějícími parametry – Kronecker`s delta.

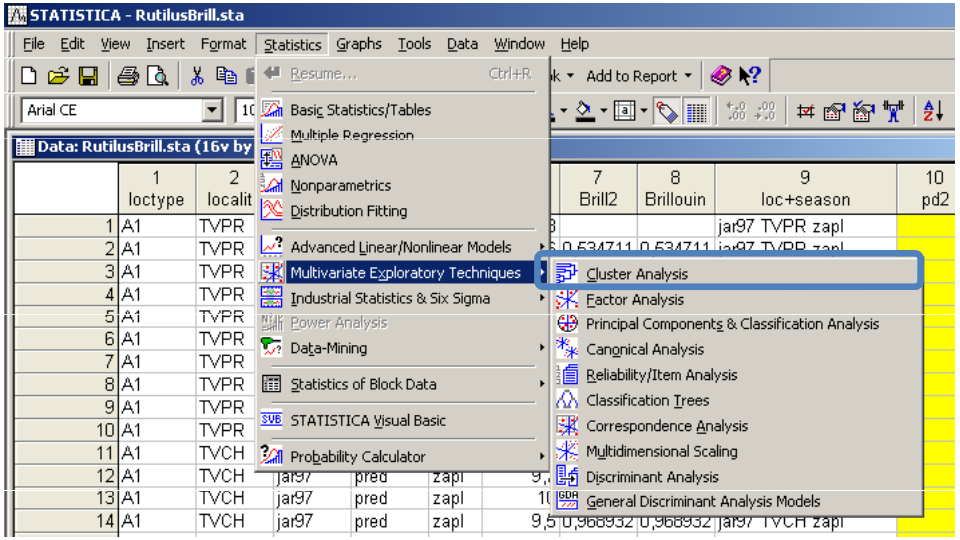
Více informací a další měření vzdáleností a podobností najdete v knize **LEGENDRE, P. & LEGENDRE, L. (1998). *Numerical ecology*. Elseviere Science BV, Amsterodam.**

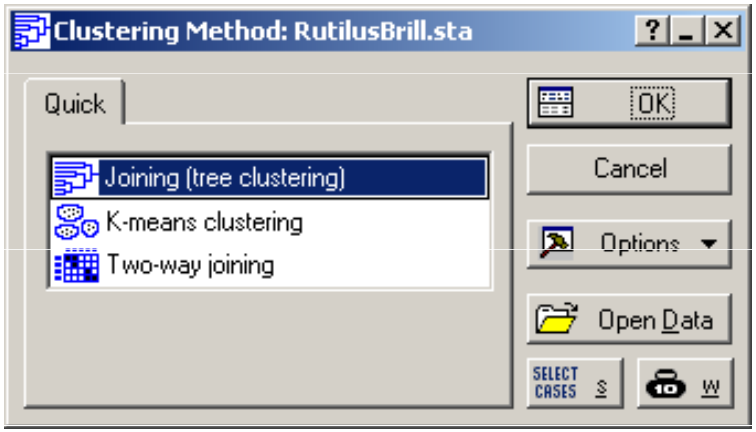
Vícerozměrné metody v STATISTICA

Statistica 9 – nabídková
větev *Multivariate Exploratory*
Techniques v menu *Statistics*



Statistics >> Multivariate Exploratory Techniques >> Cluster Analysis





- **Joining (tree clustering)** – hierarchické shlukování, podle vzdálenosti mezi objekty jsou tyto skládány do skupin pomocí různých algoritmů.
- **K – means clustering** (hypotéza existence x clusterů a její ověření analogické k ANOVA – sestavení clusterů tak aby se minimalizovala jejich vnitřní variabilita a maximalizovala variabilita mezi clustery), nehierarchické shlukování
- **Two-way joining** (shlukování je prováděno zároveň na základě jak objektů, tak parametrů)

"**Klasická" shluková analýza** hierarchicky spojující objekty do skupin podle vzdálenosti v asociační matici

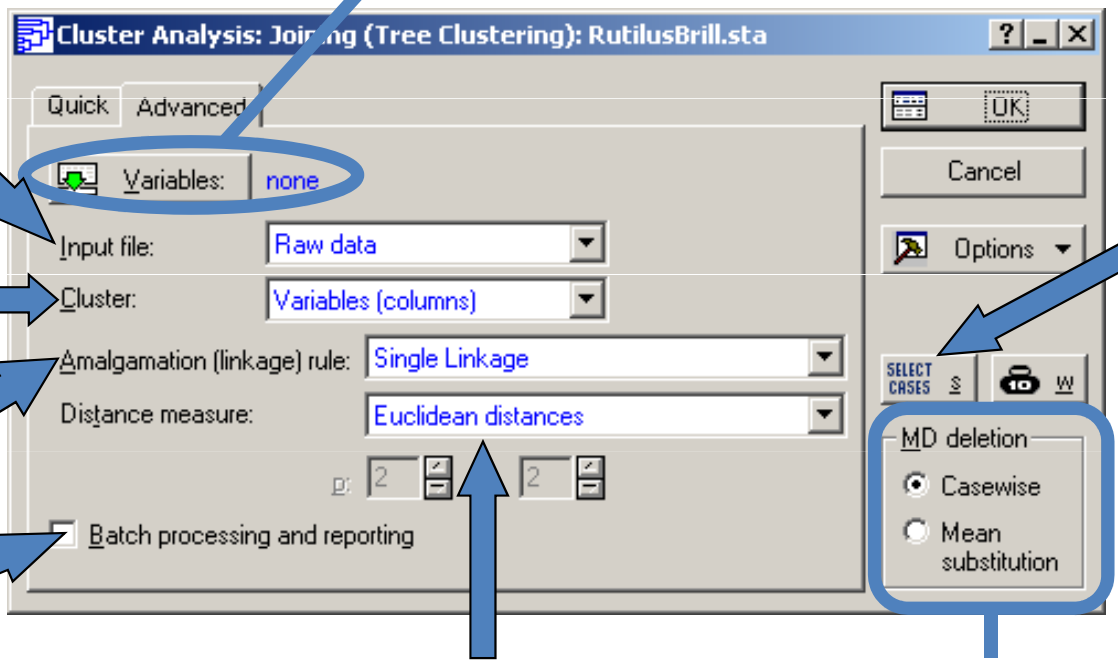
Vybrání proměnných pro výpočet

Vstupní soubor je matice objekty **x** parametry nebo matice vzdáleností

Mají být shlukovány sloupce nebo řádky vstupní matice objekty **x** parametry?

Shlukovací algoritmus

Automatizovaný výstup

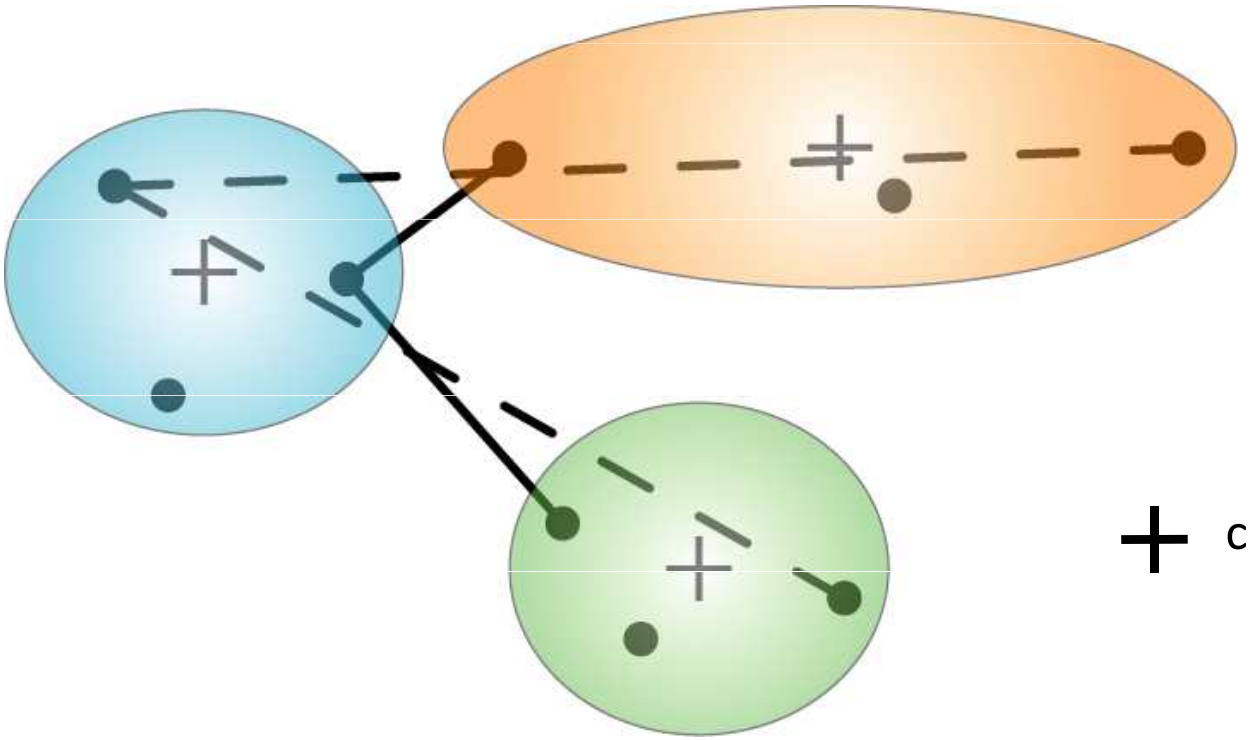


Výběr z dat

Použitá vzdálenost mezi objekty (jen matice objekty **x** parametry)

Smazání chybějících dat nebo jejich nahrazení průměrem

Joining (Tree Clustering) – shlukovací algoritmy

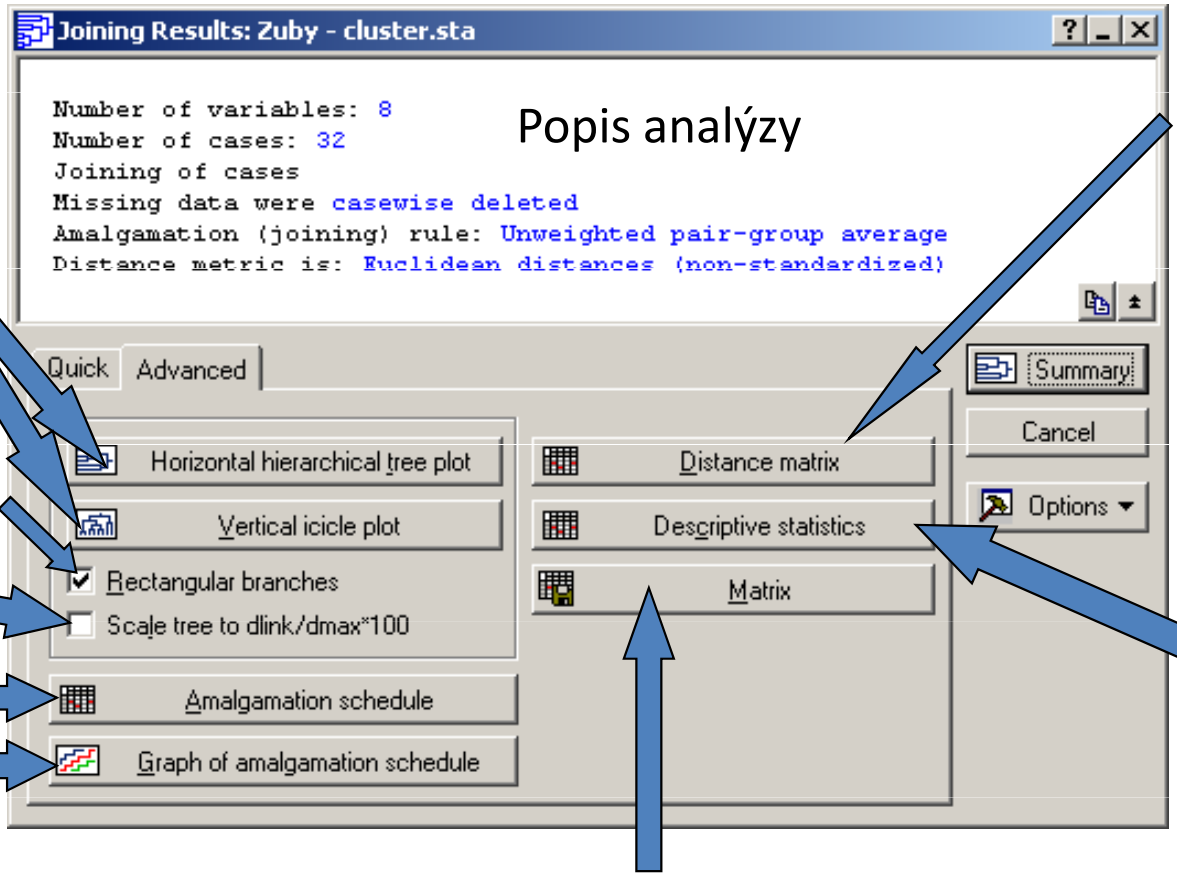


Na tuto vzdálenost se ptá **single linkage**

Na tuto vzdálenost se ptá **complete linkage**

Další metody počítají s **průměrnou vzdáleností** všech objektů shluků nebo vzdáleností **centroidů** (vzdálenost může být **vážena** velikostí shluků). **Wardova metoda** se snaží minimalizovat variabilitu uvnitř shluků.

Výsledky programu Statistica se typicky dělí na záložky **Quick** (nejdůležitější výstupy) a **Advanced** (podrobnější analýza, nastavení vlastností výstupů)



Horizontální a vertikální dendrogram

Popis analýzy

Matice vzdáleností

Pravoúhlé větve stromu

Vzdálenost v %

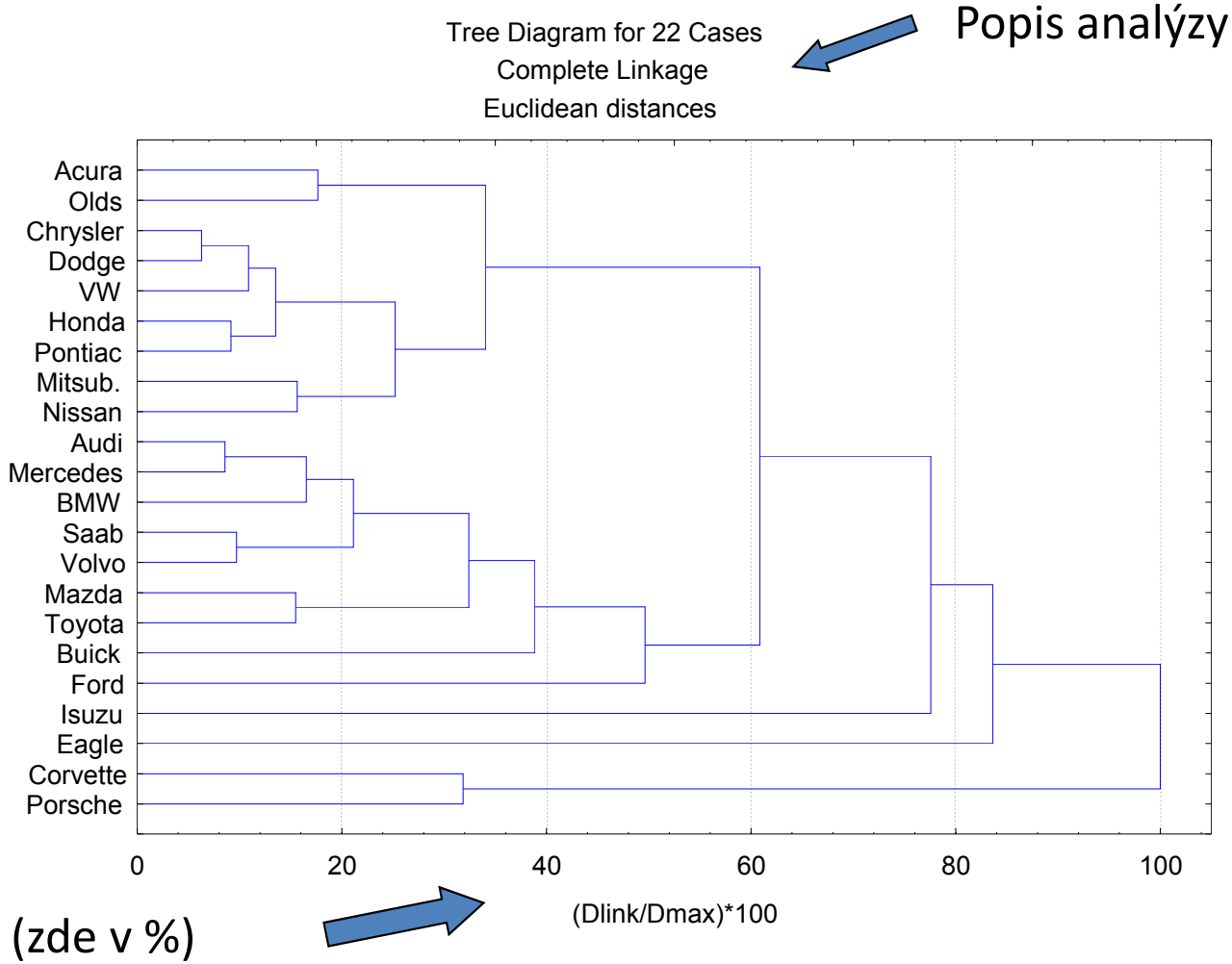
Postup skládání stromu v podobě tabulky a grafu

Popis objektů (průměr a SD)

Export matice vzdáleností (podle zvolené metriky) do speciálního souboru Statistica pro matice vzdáleností

Dendrogram představuje grafický výstup shlukové analýzy, kde jsou objekty propojeny tak, jak postupovalo jejich shlukování

Popis analýzy



Shlukované objekty

Vzdálenost (zde v %)

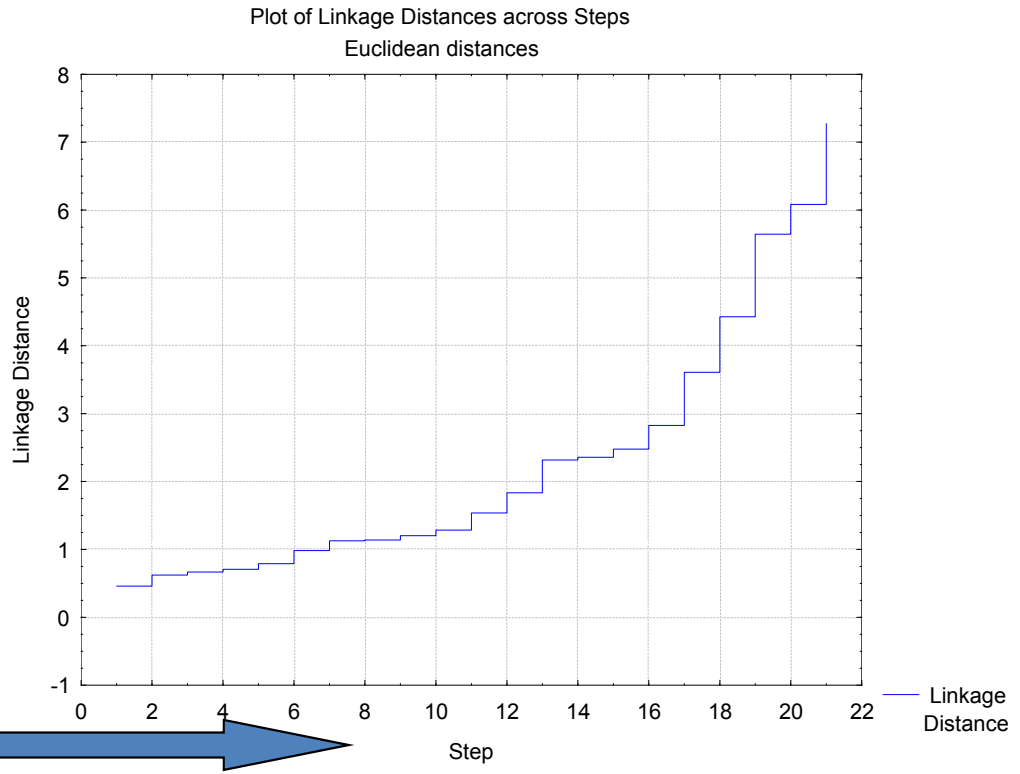
Almagenation schedule a graf poskytují uživateli přehled nad celým procesem shlukování, tj. při jaké vzdálenosti a jaké objekty nebo jejich skupiny se shlukly

Amalgamation Schedule (Cars.sta)
Complete Linkage
Euclidean distances

Linkage distance	Obj. No. 1	Obj. No. 2	Obj. No. 3	Obj. No. 4	Obj. No. 5
,4580483	Chrysler	Dodge			
,6231085	Audi	Mercedes			
,6670490	Honda	Pontiac			
,7060042	Saab	Volvo			
[illegible]	Chrysler	Dodge	VW		
[illegible]	Chrysler	Dodge	VW	Honda	
1,127473	[illegible]	Toyota			
1,137488	Mitsub.	[illegible]issan			
1,202407	Audi	Merce[illegible]	BMW		
1,284603	Acura	Olds			
1,537968	Audi	Mercedes	BM[illegible]	[illegible]aab	
1,834401	Chrysler	Dodge	VW	Hond[illegible]	

Shlukované objekty

Vzdálenost na níž došlo k shlukování



Kroky shlukování

Joining (Tree Clustering) – asociační matice

Asociační matice představují speciální typ souborů programu Statistica (přípona .smx), jde o čtvercové matice nesoucí informaci o vztazích mezi řádky a sloupci, tvoří alternativní vstup pro vícerozměrné analýzy,některé analýzy lze provádět pouze na datech v tomto formátu. Na rozdíl od běžných souborů obsahují 4 speciální řádky, pro správnou funkci je nezbytné dodržet jejich přesnou syntaxi.

Vlastní matice vzdáleností

	Var 1	Var 2	Var 3
Var 1	1.00	.20	.30
Var 2	.20	1.00	.10
Var 3	.30	.10	1.00
Means	12	11	10
Std. Dev.	3	5	2
No. Cases	50		
Matrix	1		

Průměr a SD proměnných (není nutné pro matici podobností a nepodobností)

Počet případů = počet z nějž byla matice vytvořena, ne počet jejích řádků

Typ matice 1 = korelace, 2 = podobnosti, 3 = nepodobnosti, 4 = kovariance

Shluková analýza K-means clustering

K-means clustering se snaží rozdělit objekty do zadaného počtu shluků tak, aby byla minimalizována variabilita uvnitř shluků a maximalizována mezi shluky

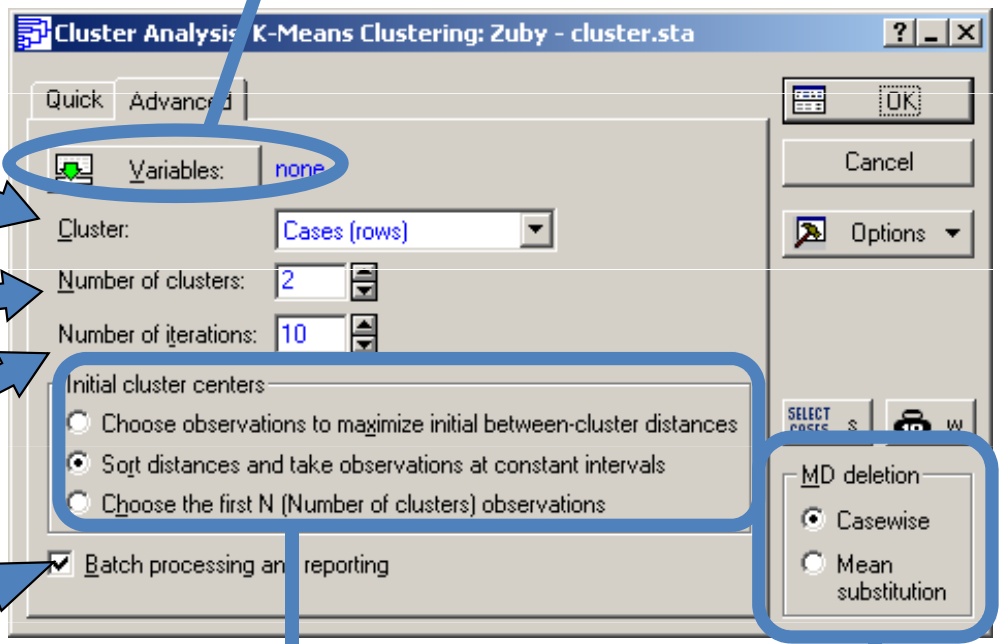
Vybrání proměnných pro výpočet

Mají být shlukovány sloupce nebo řádky vstupní matice objekty **x** parametry?

Počet očekávaných shluků

Počet iterací – kroků výpočtu

Automatizovaný výstup



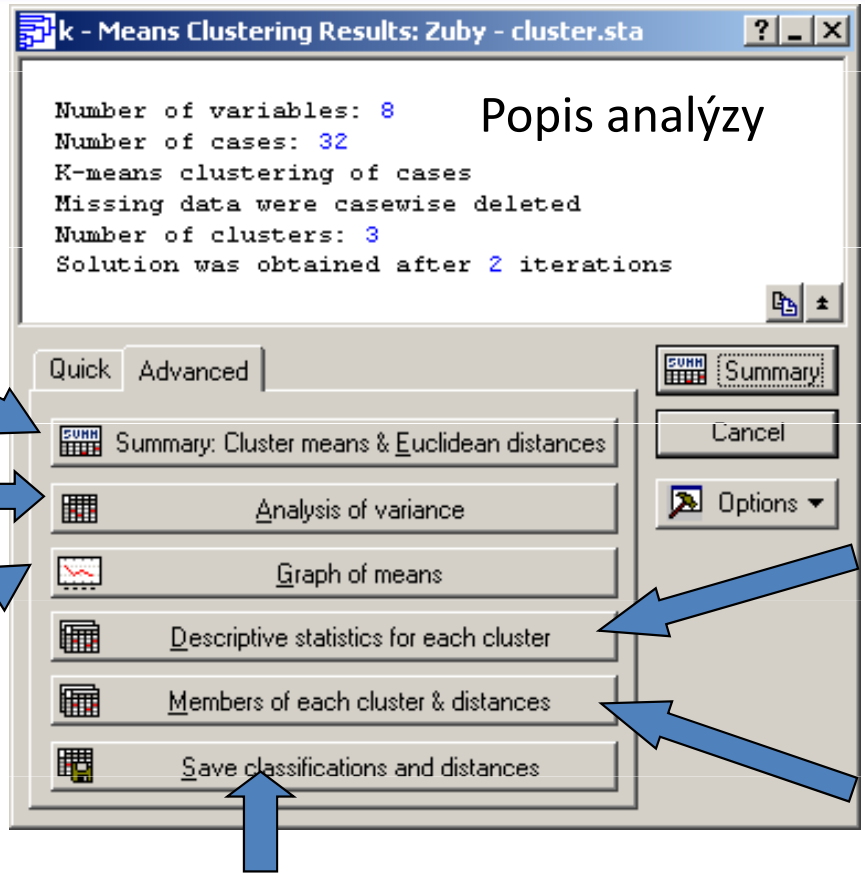
Nastavení počátečních shluků, od nichž se výpočet odvíjí

Smazání chybějících dat nebo jejich nahrazení průměrem

IBA

K-means clustering - výsledky

K-means clustering pracuje s objekty pouze na základě Euklidovské vzdálenosti, na tuto skutečnost je nezbytné pamatovat pokud tato metrika není pro data vhodná.



Euklidovská vzdálenost středu shluků

ANOVA pro jednotlivé proměnné

Graf průměrů jednotlivých proměnných v shlucích

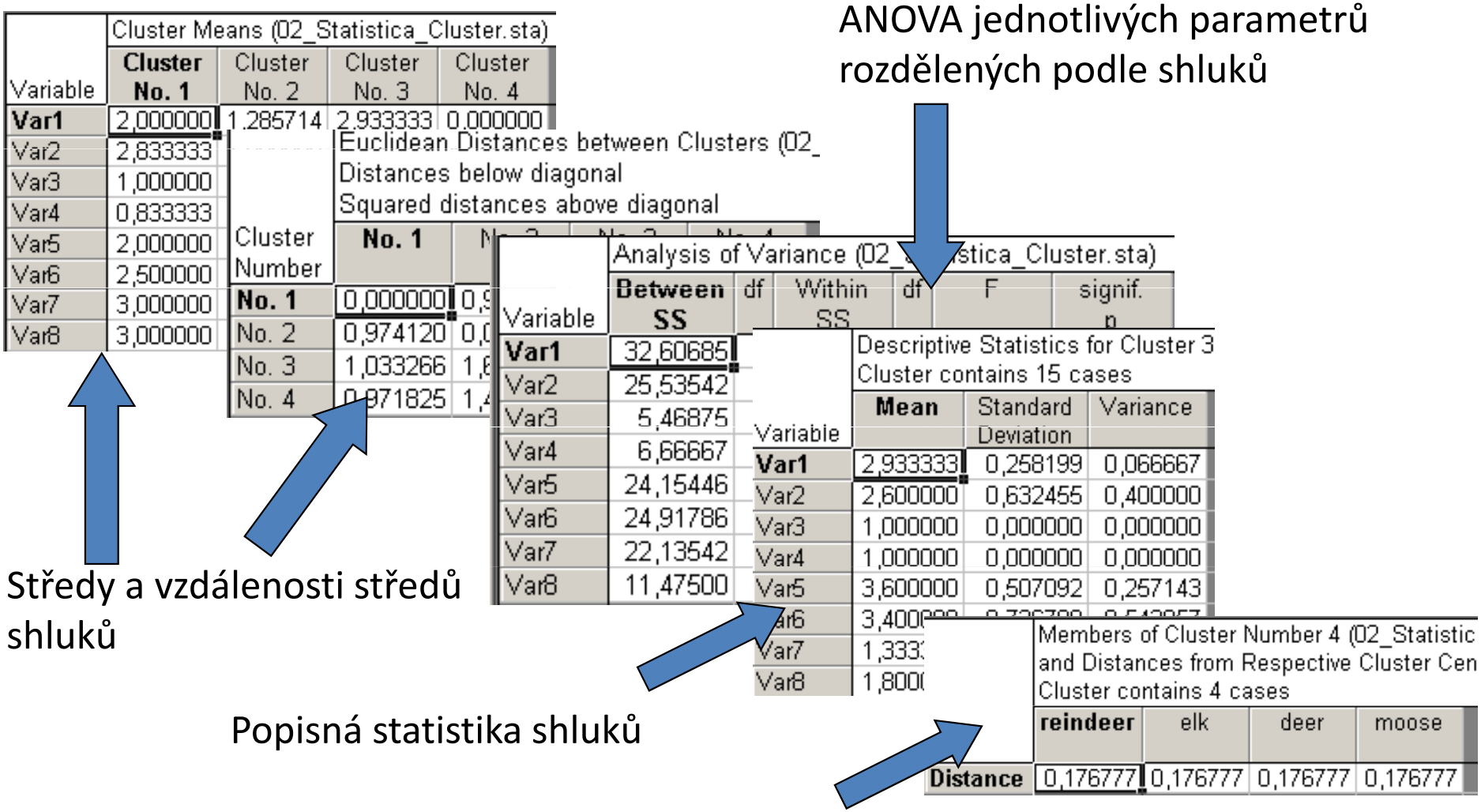
Průměr, rozptyl, SD parametrů v shlucích

Objekty v shlucích a jejich vzdálenost od centroidu

Uloží příslušnost k shluku doplněnou o vzdálenost k centroidu pro všechny objekty (+ vybrané parametry).

IBA MU

K-means clustering – tabulky výsledků



Cluster Means (02_Statistica_Cluster.sta)

Variable	Cluster No. 1	Cluster No. 2	Cluster No. 3	Cluster No. 4
Var1	2,000000	1,285714	2,933333	0,000000
Var2	2,833333			
Var3	1,000000			
Var4	0,833333			
Var5	2,000000			
Var6	2,500000			
Var7	3,000000			
Var8	3,000000			

Euclidean Distances between Clusters (02_
Distances below diagonal
Squared distances above diagonal

Cluster Number	No. 1	
No. 1	0,000000	0,9
No. 2	0,974120	0,0
No. 3	1,033266	1,8
No. 4	0,971825	1,4

Analysis of Variance (02_ stica_Cluster.sta)

Variable	Between SS	df	Within SS	df	F	signif. p
Var1	32,60685					
Var2	25,53542					
Var3	5,46875					
Var4	6,66667					
Var5	24,15446					
Var6	24,91786					
Var7	22,13542					
Var8	11,47500					

Descriptive Statistics for Cluster 3
Cluster contains 15 cases

Variable	Mean	Standard Deviation	Variance
Var1	2,933333	0,258199	0,066667
Var2	2,600000	0,632455	0,400000
Var3	1,000000	0,000000	0,000000
Var4	1,000000	0,000000	0,000000
Var5	3,600000	0,507092	0,257143
Var6	3,400		
Var7	1,333		
Var8	1,800		

Members of Cluster Number 4 (02_Statistic
and Distances from Respective Cluster Cen
Cluster contains 4 cases

	reindeer	elk	deer	moose
Distance	0,176777	0,176777	0,176777	0,176777

ANOVA jednotlivých parametrů rozdělených podle shluků

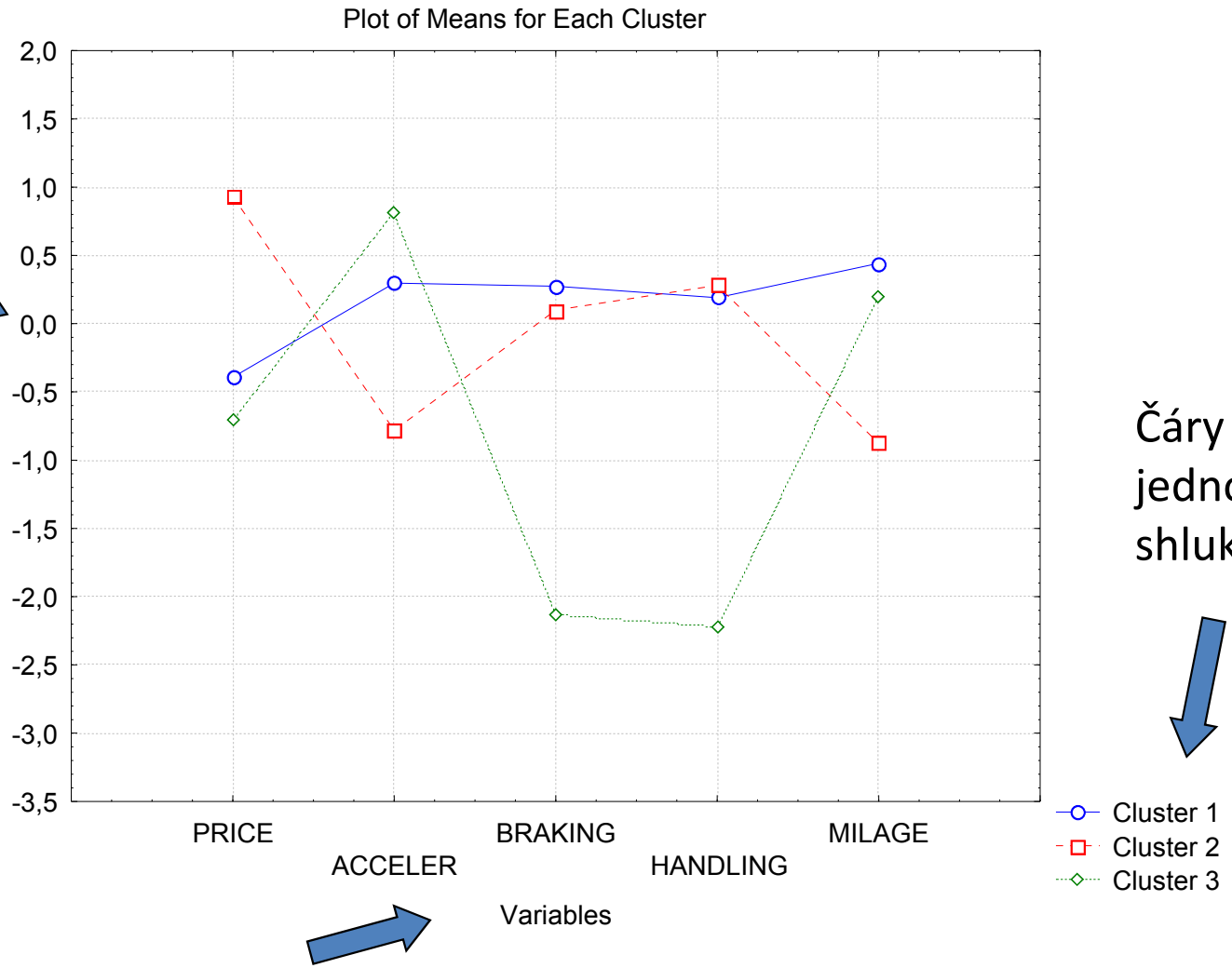
Středy a vzdálenosti středů shluků

Popisná statistika shluků

Členové shluku a jejich vzdálenost od středu shluku

K-means clustering – průměry parametrů

Průměry v shlucích



Čáry pro jednotlivé shluky

Jednotlivé parametry